IN THE CLAIMS

1. (currently amended) A method, comprising:

analyzing source codes of a main thread having one or more delinquent loads, the one or more delinquent loads representing loads which likely suffer cache misses during an execution of the main thread, the source codes including one or more code regions, each code region corresponding to a sequence of instructions representing an iteration loop in the source codes, the one or more code regions sharing at least one instruction in the source codes;

estimating a communication cost between the main thread and each code region;

selecting a code region from the one or more code regions for one or more helper threads with respect to the main thread based on the communication cost, wherein selecting the code region further comprises determining a synchronization period for the helper thread to synchronize the main thread and the helper thread, the helper thread performing its tasks within the synchronization period; and

generating codes for the one or more helper threads, the one or more helper threads being speculatively executed in parallel with the main thread to perform one or more tasks for the selected code region of the main thread.

2. (previously amended) The method of claim 1, wherein analyzing the source codes comprises:

generating one or more profiles for cache misses of the selected code region; and

analyzing the one or more profiles to identify one or more candidates for thread-based prefetch operations.

3. (previously amended) The method of claim 2, wherein generating one or more profiles comprises:

executing an application associated with the main thread with debug information; and

sampling cache misses and accumulating hardware counter for each static load of the selected code region to generate the one or more profiles for each cache hierarchy.

4. (previously presented) The method of claim 3, wherein analyzing the one or more profiles comprises:

correlating the one or more profiles with respective source code based on the debug information; and

identifying top loads that contribute cache misses above a predetermined level as the delinquent loads.

5. (previously presented) The method of claim 1, wherein analyzing the source codes comprises:

building a dependent graph that captures data and control dependencies of the main thread; and

performing slicing operations on the main thread based on the dependent graph to generate code slices, each code slice corresponding to one of the one or more delinquent loads.

6. (previously amended) The method of claim 5, wherein selecting the code region further comprises:

limiting traversal of the dependency graph to be within the selected code region for the slicing operations;

merging two or more of the code slices into a helper thread of the one or more helper threads to minimize code duplication;

computing liveness information providing communication cost between the main thread on the one of the helper threads;

determining a communication scheme communicating live-in values between the main thread and the helper thread according to the liveness information, wherein the live-in values are accessed in the helper thread without re-computation; and

determining a change in size of the helper thread according to one of the slicing operations corresponding to a separate code region of the one or more overlapping code regions,

wherein the selected code region encompasses the separate code region, wherein the change reduces the size of the helper thread.

7. (canceled)

8. (currently amended) A machine-readable storage medium having executable code to cause a machine to perform a method, the method comprising:

analyzing source codes of a main thread having one or more delinquent loads, the one or more delinquent loads representing loads which likely suffer cache misses during an execution of the main thread, the source codes including one or more code regions, each code region corresponding to a sequence of instructions representing an iteration loop in the source codes, the one or more code regions sharing at least one instruction in the source codes;

selecting a code region from the one or more code regions for one or more helper threads with respect to the main thread based on the analysis; and

generating software codes for the one or more helper threads, the one or more helper threads to be speculatively executed in parallel with the main thread to perform one or more prefetching tasks for the selected code region of the main thread,

wherein the generated software codes include synchronization codes for the one or more helper threads to synchronize with the main thread during the execution,

wherein synchronization codes are generated based on a synchronization period for the one or more helper threads to synchronize with each other during the execution, and wherein the synchronization period is based on a distance between the one or more helper threads and the main thread.

9. (previously amended) The machine-readable storage medium of claim 8, wherein analyzing the source codes comprises:

generating one or more profiles for cache misses of the selected code region; and

analyzing the one or more profiles to identify one or more candidates for thread-based prefetch operations.

10. (previously amended) The machine-readable storage medium of claim 9, wherein generating one or more profiles comprises:

executing an application associated with the main thread with debug information; and

sampling cache misses and accumulating hardware counter for each static load of the selected code region to generate the one or more profiles for each cache hierarchy.

11. (previously presented) The machine-readable storage medium of claim 10, wherein analyzing the one or more profiles comprises:

correlating the one or more profiles with respective source code based on the debug information; and

identifying top loads that contribute cache misses above a predetermined level as the delinquent loads.

12. (previously presented) The machine-readable storage medium of claim 8, wherein analyzing the source codes comprises:

building a dependent graph that captures data and control dependencies of the main thread; and

performing slicing operations on the main thread based on the dependent graph to generate code slices, each code slice corresponding to one of the one or more delinquent loads.

13. (previously amended) The machine-readable storage medium of claim 12, wherein selecting the code region further comprises:

limiting traversal of the dependency graph to be within the selected code region for the slicing operations;

merging two or more of the code slices into a helper thread of the one or more helper threads to minimize code duplication;

computing liveness information providing communication cost between the main thread on the one of the helper threads;

determining a communication scheme communicating live-in values between the main thread and the helper thread according to the liveness information, wherein the live-in values are accessed in the helper thread without re-computation; and

determining a change in size of the helper thread according to one of the slicing operations corresponding to a separate code region of the one or more overlapping code regions, wherein the selected code region encompasses the separate code region, wherein the change reduces the size of the helper thread.

14. (previously amended) The machine-readable storage medium of claim 8, wherein generating the software codes further comprises determining a synchronization period for the software codes such that the helper thread is to perform its respective prefetching tasks within the synchronization period.

15. (currently amended) A data processing system, comprising:

a memory storing executable instructions including a compiler; and

a processor coupled to the memory to execute the executable instructions from the memory for performing multi threading operations, the processor being configured to

analyze source codes of a main thread having one or more delinquent loads, the one or more delinquent loads representing loads which likely suffer cache misses during an execution of the main thread, the source codes including one or more code regions, each code region corresponding to a sequence of instructions representing an iteration loop in the source codes, the one or more code regions sharing at least one instruction in the source codes,

estimate a communication cost between the main thread and each code region,

select the code region from the one or more code regions for one or more helper threads with respect to the main thread based on the communication cost, wherein selecting the code region further comprises determining a synchronization period for the helper thread to synchronize the main thread and the helper thread, the helper thread performing its tasks within the synchronization period, and

generate code for the one or more helper threads, the one or more helper threads being speculatively executed in parallel with the main thread to perform one or more tasks for the selected code region of the main thread.

16. (original) The data processing system of claim 15, wherein the process is executed by a compiler during a compilation of an application.

17. (currently amended) A method, comprising:

executing a main thread of an application in a multi-threading system; and

spawning one or more helper threads from the main thread created from source codes including one or more code regions sharing at least one instruction in the source codes, each code region associated with an estimation of communication cost with the main thread, the code region corresponding to a sequence of instructions representing an iteration loop in the source codes,

the one or more helper thread to perform one or more computations for the main thread when the main thread enters a code region selected from the one or more code regions based on the estimation of communication cost, the selected code region having one or more delinquent loads,

the one or more helper threads being created separately from the source codes of the main thread during a compilation of the source codes for the main thread,

wherein the one or more helper threads are synchronized with the main thread based on a synchronization period determined for the one or more helper threads, so that the one or more helper threads perform the one or more computations during the synchronization period.

18. (original) The method of claim 17, further comprising:

creating a thread pool to maintain a list of thread contexts; and

allocating one or more thread contexts from the thread pool to generate the one or more helper threads.

19. (previously presented) The method of claim 18, further comprising:

terminating the one or more helper threads when the main thread exits the code region; and

releasing the thread contexts associated with the one or more helper threads back to the thread pool.

20. (currently amended) The method of claim 17, wherein the one or more help threads are placed in a run queue prior to execution, further comprising

determining a time period for each of the helper threads in the run queue, each of the helper threads being terminated from the run queue when the respective time period expires, thereby releasing resources back to the main thread.

21. (original) The method of claim 20, wherein each of the helper threads terminates when the time period expires even if the respective helper thread has not been accessed by the main thread.

22. (previously presented) The method of claim 17, further comprising discarding results generated by the one or more helper threads when the main thread exits the code region, the results not being reused by another code region of the main thread.

23. (currently amended) A machine-readable storage medium having executable code to cause a machine to perform a method, the method comprising:

executing a main thread of an application in a multi-threading system; and

spawning one or more helper threads from the main thread created from source codes including one or more code regions sharing at least one instruction of the source codes, each code region associated with an estimation of communication cost with the main thread, the code region corresponding to a sequence of instructions representing an iteration loop in the source codes,

the one or more helper threads to perform one or more computations for the main thread when the main thread enters a code region selected from the one or more code regions based on the estimation of communication cost, the selected code region having one or more delinquent loads, the one or more helper thread being created separately from the source codes of the main thread during a compilation of the source codes for the main thread,

wherein the one or more helper threads are synchronized with the main thread based on a synchronization period determined for the one or more helper threads, so that the one or more helper threads perform the one or more computations during the synchronization period.

24. (previously presented) The machine-readable storage medium of claim 23, wherein the method further comprises:

creating a thread pool to maintain a list of thread contexts; and

allocating one or more thread contexts from the thread pool to generate the one or more helper threads.

25. (previously presented) The machine-readable storage medium of claim 24, wherein the method further comprises:

terminating the one or more helper threads when the main thread exits the code region; and

releasing the thread contexts associated with the one or more helper threads back to the thread pool.

26. (currently amended) The machine-readable storage medium of claim 23, wherein the one or more help threads are placed in a run queue prior to execution, and wherein the method further comprises determining a time period for each of the helper threads in the run queue, each of the helper threads being terminated from the run queue when the respective time period expires, thereby releasing resources back to the main thread.

27. (previously presented) The machine-readable storage medium of claim 26, wherein each of the helper threads terminates when the time period expires even if the respective helper thread has not been accessed by the main thread.

28. (previously presented) The machine-readable storage medium of claim 23, wherein the method further comprises discarding results generated by the one or more helper threads when the main thread exits the code region, the results not being reused by another code region of the main thread.

29. (currently amended) A data processing system, comprising:

a memory storing executable instructions for multi-threading operations; and

a processor coupled to the memory to execute the executable instructions from the memory, the processor being configured to

execute a main thread of an application in a multi-threading system, and

spawn one or more helper threads from the main thread created from source codes including one or more code regions sharing at least one instruction of the source codes, each code region associated with an estimation of communication cost with the main thread, the code

region corresponding to a sequence of instructions representing an iteration loop in the source codes, the one or more helper threads to perform one or more computations for the main thread when the main thread enters a code region selected from the one or more code regions based on the estimation of communication cost, the selected code region having one or more delinquent loads,

the one or more helper thread being created separately from the source codes of the main thread during a compilation of the source codes for the main thread,

wherein the one or more helper threads are synchronized with the main thread based on a synchronization period determined for the one or more helper threads, so that the one or more helper threads perform the one or more computations during the synchronization period.

30. (Canceled)